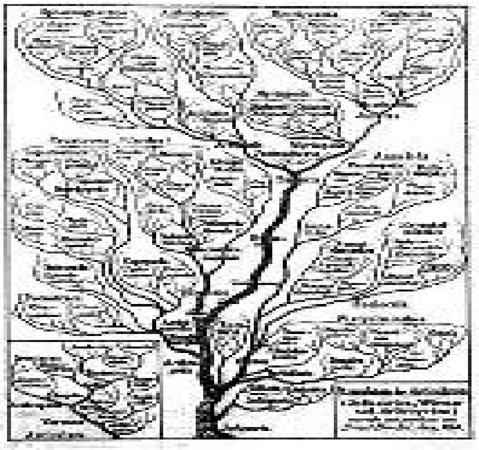
Computational methods in phylogenetic analysis

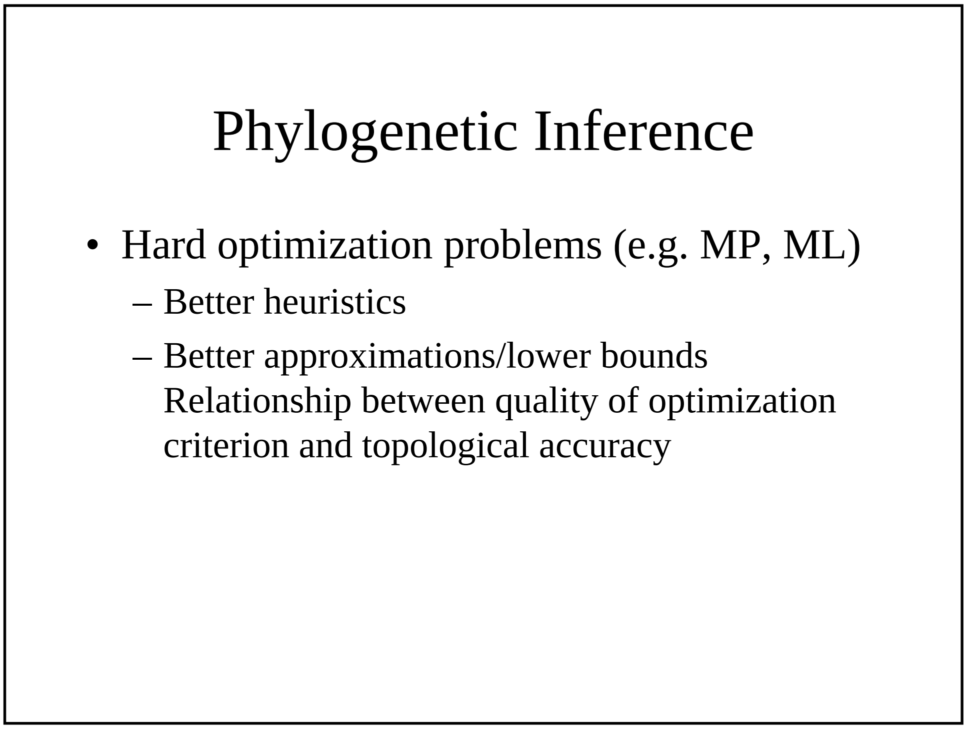
Tutorial at CSB 2004

Tandy Warnow

Reconstructing the "Tree" of Life

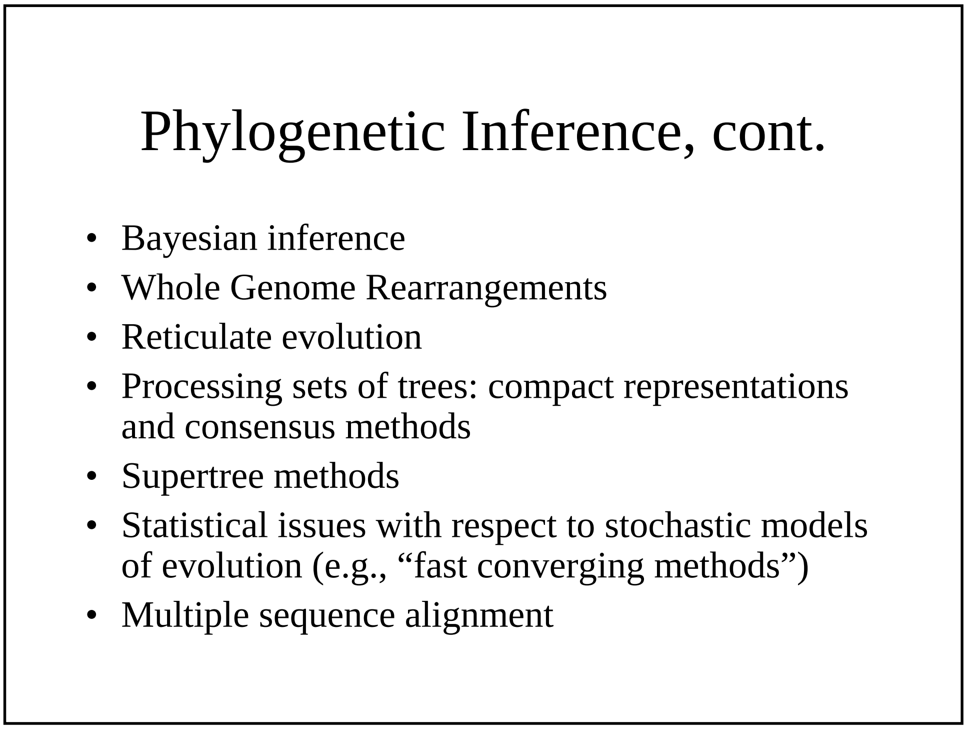


Handling large datasets: millions of species



Phylogenetic Inference

- Hard optimization problems (e.g. MP, ML)
  - – Better heuristics
  - – Better approximations/lower bounds Relationship between quality of optimization criterion and topological accuracy



Phylogenetic Inference, cont.

- Bayesian inference
- Whole Genome Rearrangements
- Reticulate evolution
- Processing sets of trees: compact representations and consensus methods
- Supertree methods
- Statistical issues with respect to stochastic models of evolution (e.g., “fast converging methods”)
- Multiple sequence alignment

Major challenge: MP and ML

- Maximum Parsimony (MP) and Maximum Likelihood (ML) remain the methods of choice for most systematists
- The main challenge here is to make it possible to obtain good solutions to MP or ML in reasonable time periods on large datasets

Outline

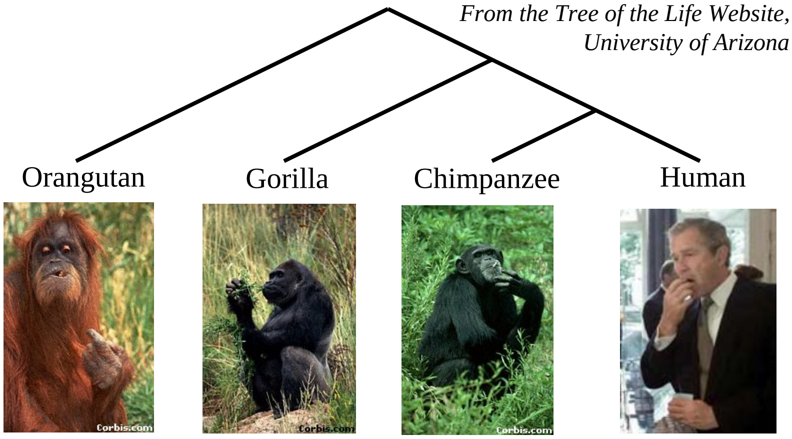
- Part I (Basics): 40 minutes
- Part II (Models of evolution): 20 min.
- Part III (Distance-based methods): 30 min.
- Part IV (Maximum Parsimony): 30 min.
- Part V (Maximum Likelihood): 15 minutes
- Part VI (Open problems/research directions): 30 minutes

Part I: Basics (40 minutes)

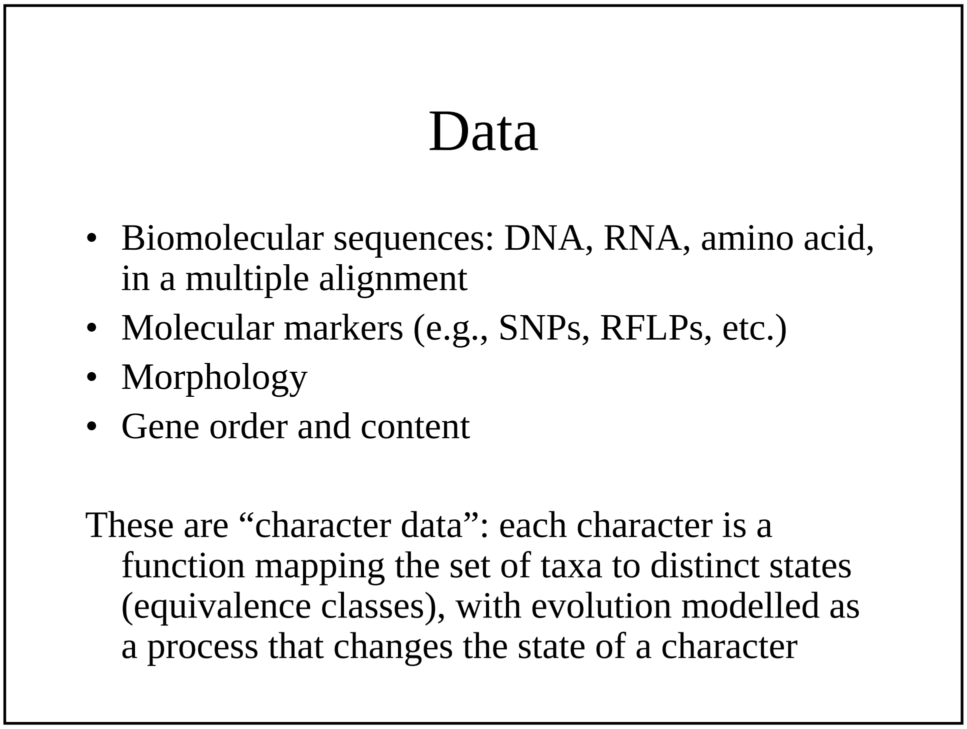
Questions:

- What is a phylogeny?
- What data are used?
- What are the most popular methods?
- What is meant by “accuracy”, and how is it measured?
- What is involved in a phylogenetic analysis?

Phylogeny



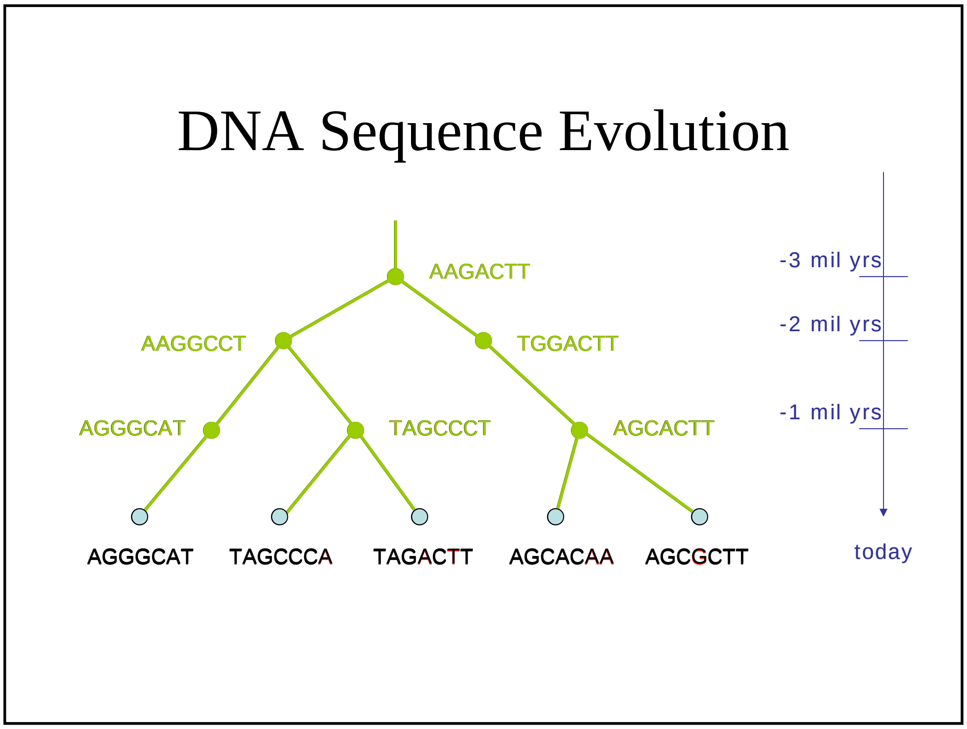
From the Tree of the Life Website, University of Arizona



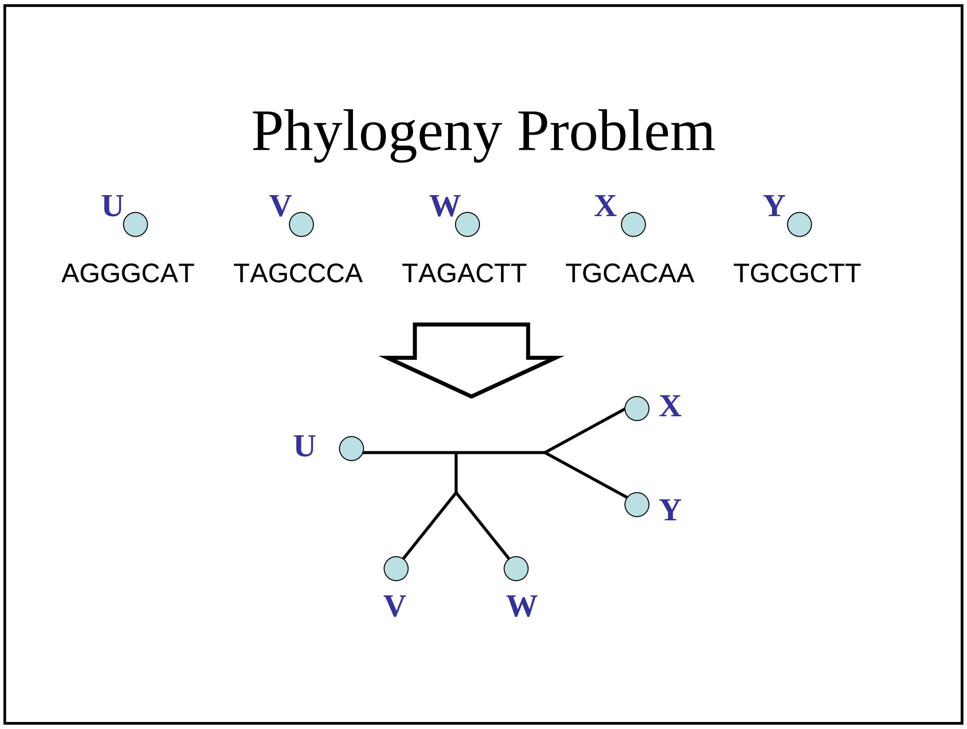
Data

- Biomolecular sequences: DNA, RNA, amino acid, in a multiple alignment
- Molecular markers (e.g., SNPs, RFLPs, etc.)
- Morphology
- Gene order and content

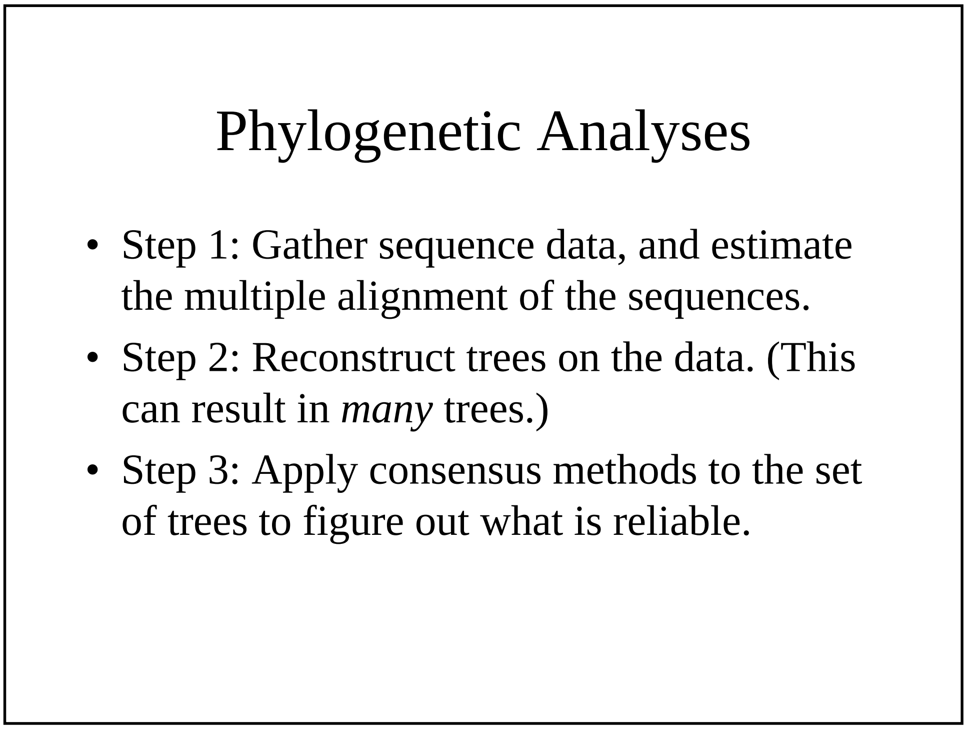
These are “character data”: each character is a function mapping the set of taxa to distinct states (equivalence classes), with evolution modelled as a process that changes the state of a character



DNA Sequence Evolution



Phylogeny Problem



Phylogenetic Analyses

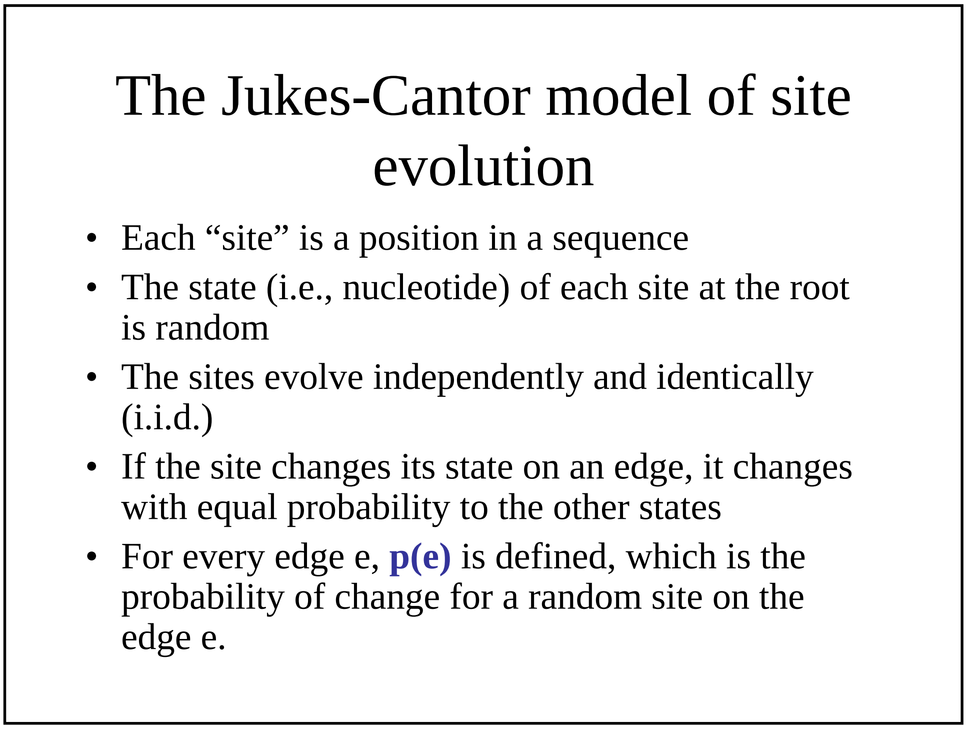
- Step 1: Gather sequence data, and estimate the multiple alignment of the sequences.
- Step 2: Reconstruct trees on the data. (This can result in *many* trees.)
- Step 3: Apply consensus methods to the set of trees to figure out what is reliable.

Reconstruction methods

- Much software exists, most of which attempt to solve one of two major optimization criteria: Maximum Parsimony and Maximum Likelihood. The most frequently used software package is PAUP*, which contains many different heuristics.
- Methods for phylogeny reconstruction are evaluated primarily in simulation studies, based upon stochastic models of evolution.

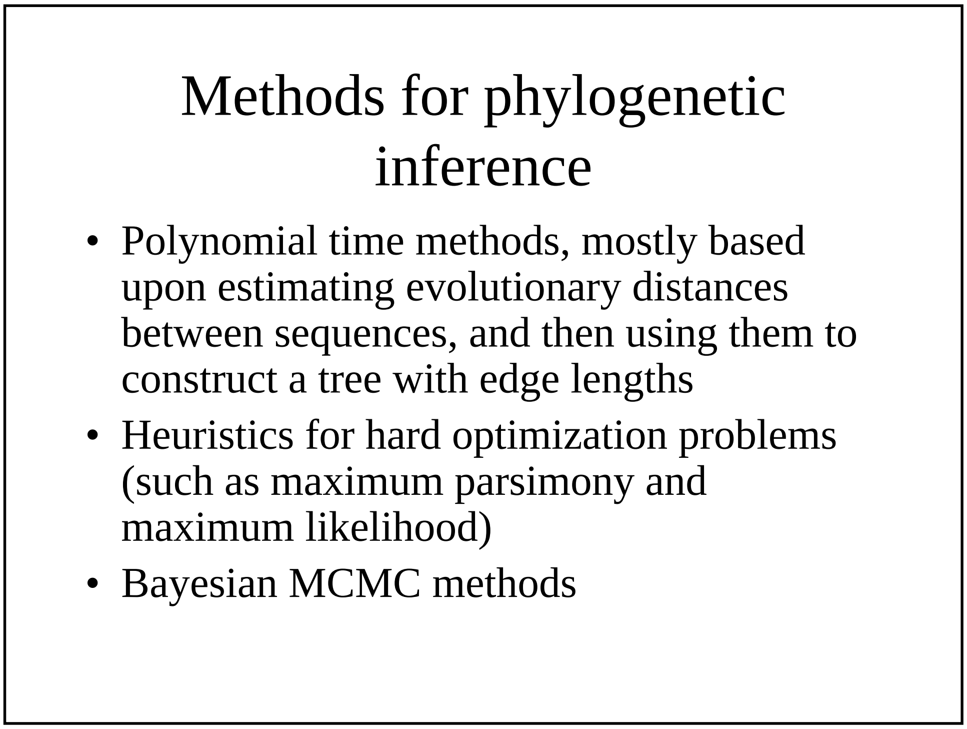
Consensus and agreement methods

- Consensus methods take a set of trees on the same set of taxa, and return a single tree on the full set. Standard approaches: strict consensus and majority tree.
- Agreement methods take a set of trees on the same set of taxa, and return a single tree on a subset of the taxa. Standard approaches: maximum agreement subtree.
- Much new research needs to be done



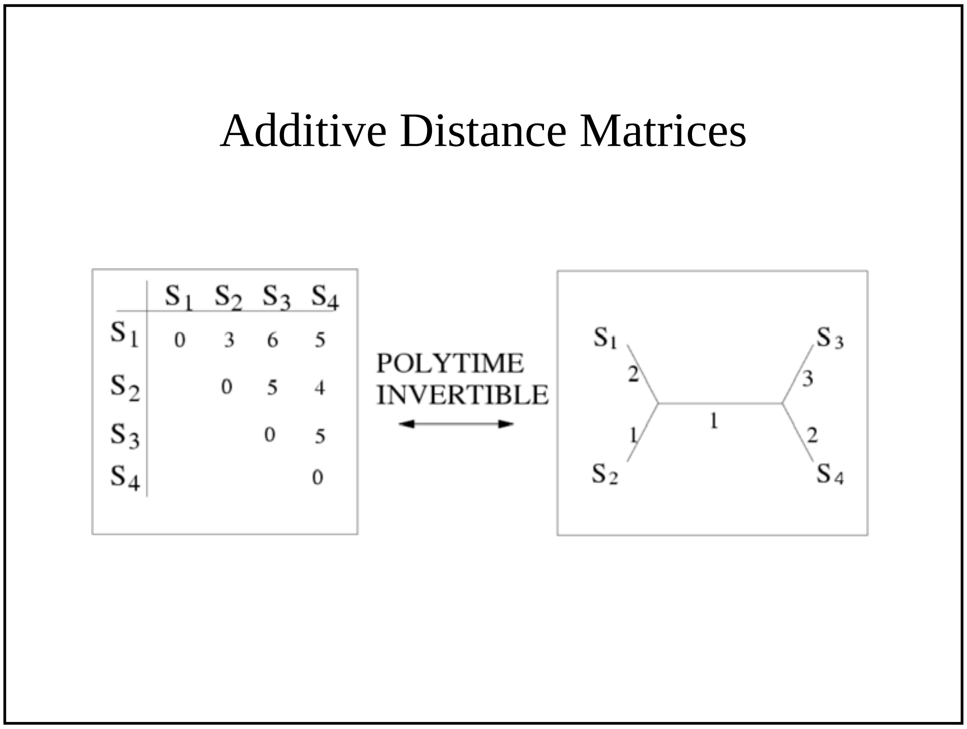
The Jukes-Cantor model of site evolution

- Each “site” is a position in a sequence
- The state (i.e., nucleotide) of each site at the root is random
- The sites evolve independently and identically (i.i.d.)
- If the site changes its state on an edge, it changes with equal probability to the other states
- For every edge e, **p(e)** is defined, which is the probability of change for a random site on the edge e.



Methods for phylogenetic inference

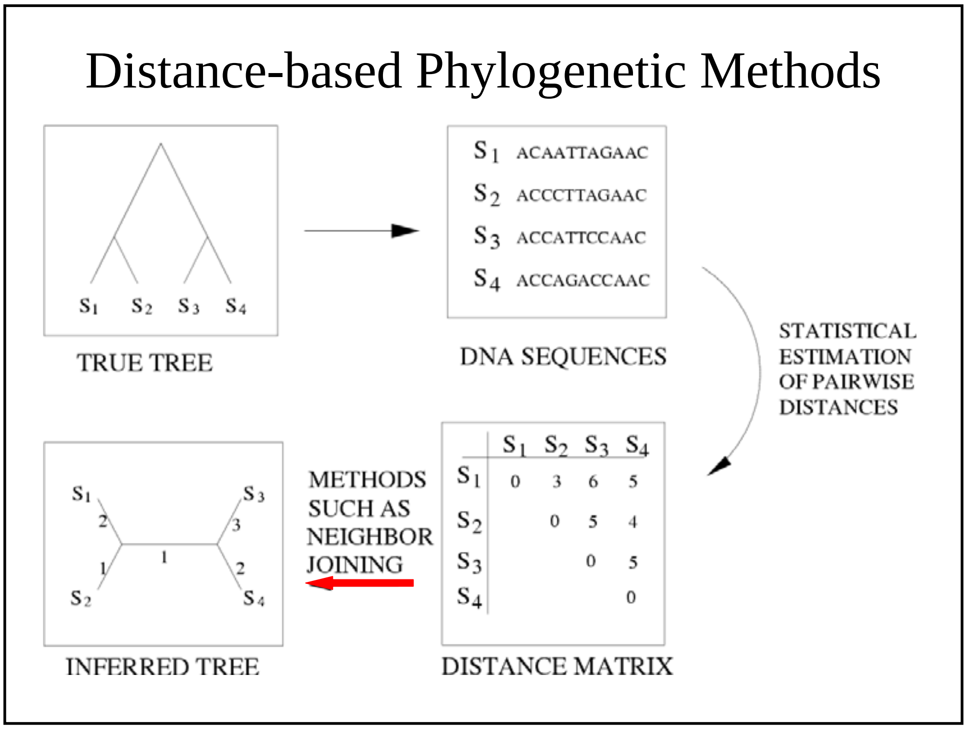
- Polynomial time methods, mostly based upon estimating evolutionary distances between sequences, and then using them to construct a tree with edge lengths
- Heuristics for hard optimization problems (such as maximum parsimony and maximum likelihood)
- Bayesian MCMC methods



Additive Distance Matrices

	S_1	S_2	S_3	S_4
S_1	0	3	6	5
S_2		0	5	4
S_3			0	5
S_4				0

POLYTIME
INVERTIBLE



Distance-based Phylogenetic Methods

TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC

DNA SEQUENCES

STATISTICAL
ESTIMATION
OF PAIRWISE
DISTANCES

	S_1	S_2	S_3	S_4
S_1	0	3	6	5
S_2		0	5	4
S_3			0	5
S_4				0

DISTANCE MATRIX

METHODS
SUCH AS
NEIGHBOR
JOINING

INFERRED TREE

Standard problem: Maximum Parsimony (Hamming distance Steiner Tree)

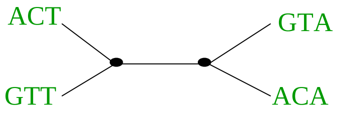
- **Input**: Set *S* of *n* aligned sequences of length k
- **Output**: A phylogenetic tree *T*
  - leaf-labeled by sequences in *S*
  - additional sequences of length *k* labeling the internal nodes of *T*

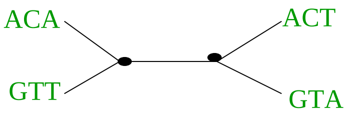
such that $\sum_{(i,j)\in E(T)} H(i,j)$ is minimized.

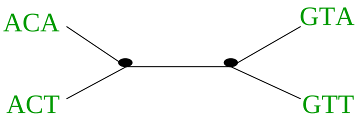
Maximum parsimony (example)

- **Input**: Four sequences
  - ACT
  - ACA
  - GTT
  - GTA
- **Question**: which of the three trees has the best MP scores?

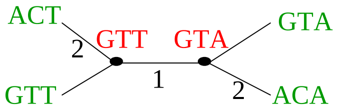
Maximum Parsimony



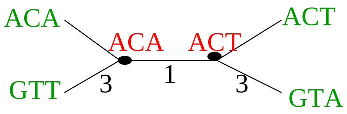




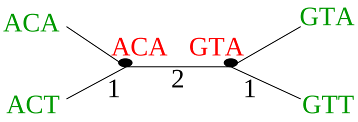
Maximum Parsimony



MP score = 5



MP score = 7

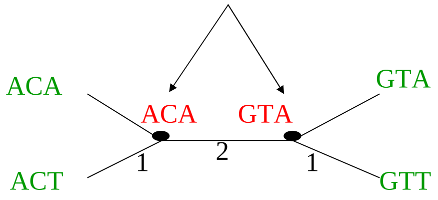


MP score = 4

Optimal MP tree

Maximum Parsimony: computational complexity

Optimal labeling can be computed in linear time O(nk)



MP score = 4

Finding the optimal MP tree is **NP-hard**

Maximum Likelihood (ML)

- Given: stochastic model of sequence evolution (e.g. Jukes-Cantor) and a set S of sequences
- Objective: Find tree T and probabilities p(e) of substitution on each edge, to maximize the probability of the data.

Preferred by some systematists, but even harder than MP in practice.

Bayesian MCMC

- Assumes a model of evolution (e.g., Jukes-Cantor)
- The basic algorithmic approach is a random walk through the space of model trees, with the probability of the data on the model tree determining whether the proposed new model tree is accepted or rejected.
- Statistics on the set of trees visited after “burn-in” constitute the output.

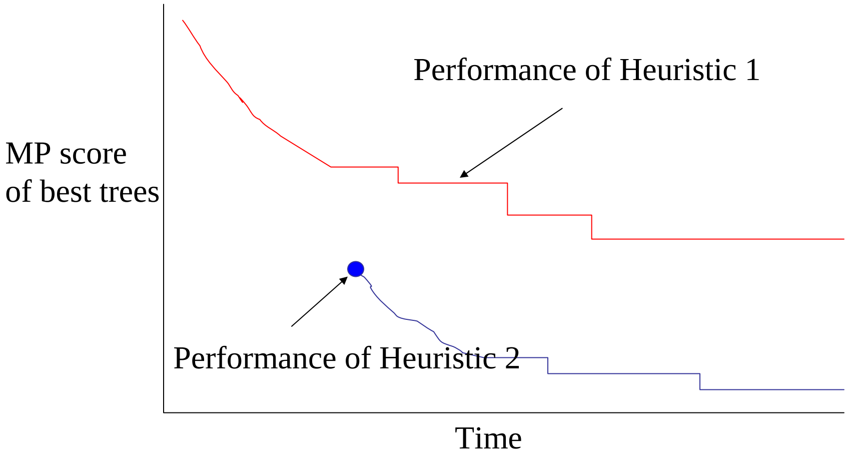
Performance criteria for phylogeny reconstruction methods

- Speed
- Space
- Optimality criterion accuracy
- “Topological accuracy” (specifically statistical consistency, convergence rate, and performance on finite data)

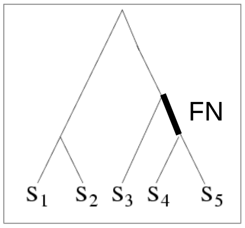
These criteria can be evaluated on real or simulated data.

Evaluating MP heuristics with respect to MP scores

Fake study



Quantifying Topological Error



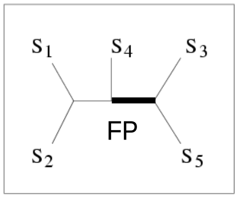
TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC
S_5	ACCAGACCGGA

DNA SEQUENCES

FN: false negative
(missing edge)
FP: false positive
(incorrect edge)

50% error rate



INFERRED TREE

Statistical performance issues

- Statistical consistency: an estimation method is statistically consistent under a model if the probability that the method returns the true tree goes to 1 as the sequence length goes to infinity
- Convergence rate: the amount of data that a method needs to return the true tree with high probability, as a function of the model tree

Practice

- In practice, most systematic biologists use either MP or ML on small datasets, and MP or MCMC methods on moderate to large datasets
- Distance-based methods (such as neighbor joining) are used by some, but are not considered as reliable as these other approaches.

Major challenges

- The main challenge here is to make it possible to obtain good solutions to MP or ML in reasonable time periods on large datasets
- MCMC methods are increasingly used (often as a surrogate for a decent ML analysis), but it is not clear how to evaluate MCMC methods

Part II: Models of evolution (20 minutes)

- Site evolution models
- Variation across sites
- Statistical performance issues: statistical identifiability, statistical consistency, convergence rates
- Special issues: molecular clock, no-common-mechanism

The Jukes-Cantor model of site evolution

- Each “site” is a position in a sequence
- The state (i.e., nucleotide) of each site at the root is random
- The sites evolve independently and identically (i.i.d.)
- If the site changes its state on an edge, it changes with equal probability to the other states
- For every edge e, **p(e)** is defined, which is the probability of change for a random site on the edge e.

General Markov (GM) Model

- A GM model tree is a pair $(T, \mathcal{M})$ where
  - T is a rooted binary tree.
  - $\mathcal{M} = \{M(e) : e \in E(T)\}$, and $M(e)$ is a stochastic substitution matrix with $\det(M(e)) \neq 0, \quad \pm 1$
  - The state at the root of T is random.
- GM contains models like Jukes-Cantor (JC), Kimura 2-Parameter (K2P), and the Generalized Time Reversible (GTR) models.

Variation across sites

- Standard assumption of how sites can vary is that each site has a multiplicative scaling factor
- Typically these scaling factors are drawn from a Gamma distribution (or Gamma plus invariant)

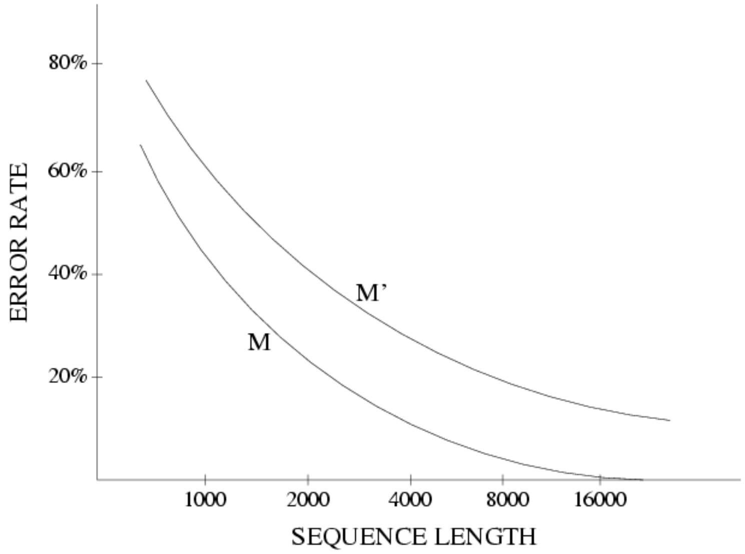
Special issues

- Molecular clock: the expected number of changes for a site is proportional to time
- No-common-mechanism model: there is a random variable for every combination of edge and site

Statistical performance issues

- Statistical consistency: an estimation method is statistically consistent under a model if the probability that the method returns the true tree goes to 1 as the sequence length goes to infinity
- Convergence rate: the amount of data that a method needs to return the true tree with high probability, as a function of the model tree

Statistical consistency and convergence rates



Statistical performance

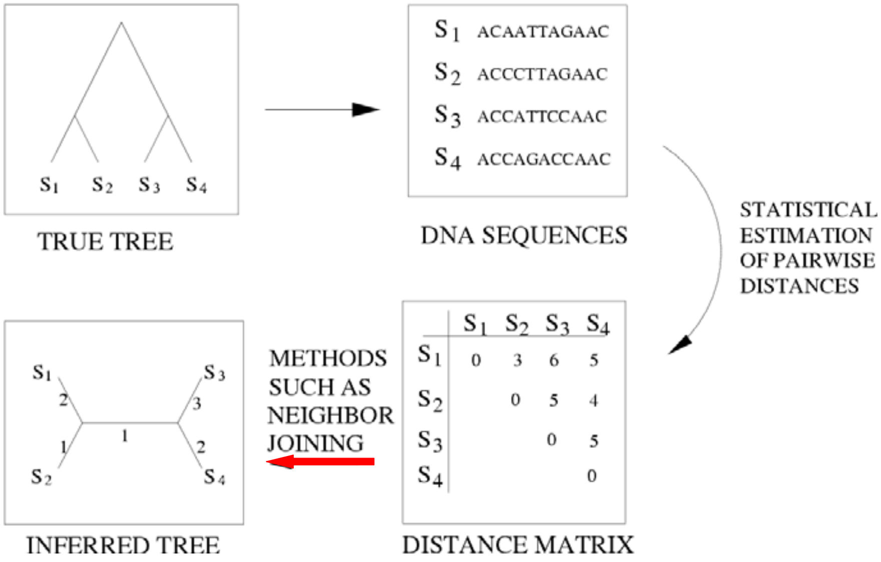
- Standard distance-based methods and Maximum Likelihood (solved exactly) are statistically consistent under the General Markov model
- Maximum Parsimony is not always statistically consistent, even for the (simplest) Jukes-Cantor model
- No method can be statistically consistent under the No Common Mechanism model - because the model is not identifiable. (In fact, under this model, MP = ML)

Part III: Distance-based methods (30 minutes)

Overview

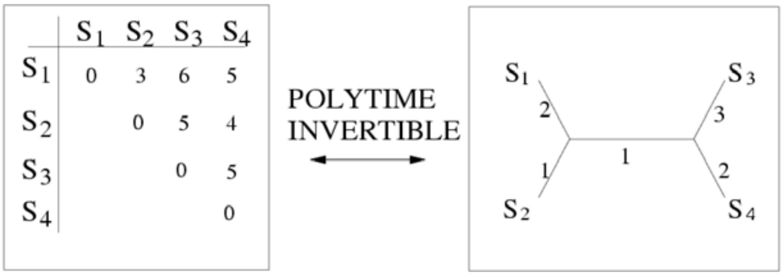
- Additive matrices and the four-point condition and method
- The Naïve Quartet Method
- Statistical consistency
- Convergence rates (sequence length requirements)
- Absolute fast convergence versus exponential convergence

Distance-based Phylogenetic Methods



	S_1	S_2	S_3	S_4
S_1	0	3	6	5
S_2		0	5	4
S_3			0	5
S_4				0

Additive Distance Matrices



	S_1	S_2	S_3	S_4
S_1	0	3	6	5
S_2		0	5	4
S_3			0	5
S_4				0

Four-point condition

- A matrix D is additive if and only if for every four indices i,j,k,l, the maximum and median of the three pairwise sums are identical

$$D_{ij}+D_{kl} < D_{ik}+D_{jl} = D_{il}+D_{jk}$$

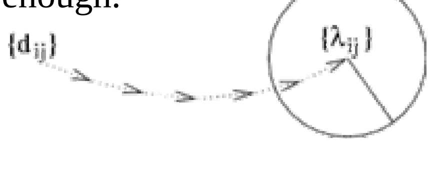
The Four-Point Method computes trees on quartets using the Four-point condition

Naïve Quartet Method

- Compute the tree on each quartet using the four-point condition
- Merge them into a tree on the entire set if they are compatible:
  - Find a sibling pair A,B
  - Recurse on S-{A}
  - If S-{A} has a tree T, insert A into T by making A a sibling to B, and return the tree

Statistical Consistency

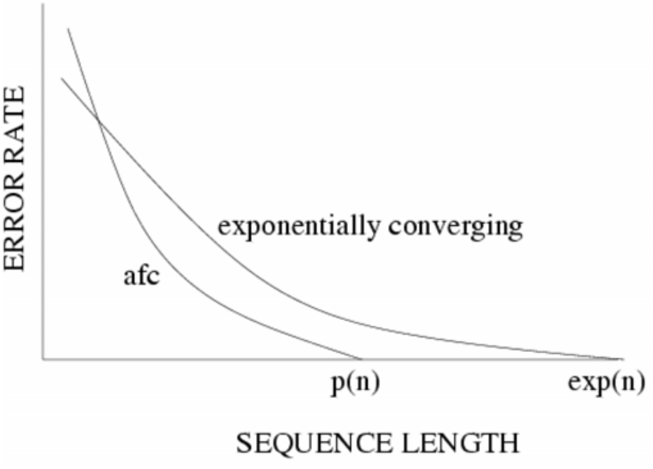
The Naïve Quartet Method (NQM) returns the true tree if $L_\infty(d,\lambda)$ is small enough.



Sequence length $\rightarrow\infty$

Hence NQM is statistically consistent for many models of evolution.
(The same result holds for many distance-based methods.)

Absolute fast convergence vs. exponential convergence



Absolute Fast Convergence

- Let $f, g \geq 0$. Define $\lambda(e) = -\log|\det(M_e)|$. We parameterize the GM model:

$$GM_{f,g} = \{(T, \mathcal{M}) \in GM : \forall e \in E(T), f \leq \lambda(e) \leq g\}$$

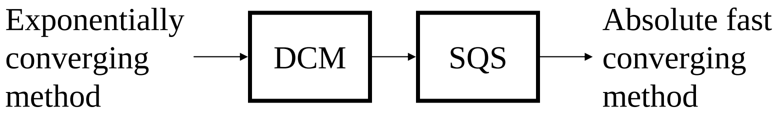
- A phylogenetic reconstruction method Φ is **absolute fast-converging (AFC)** for the GM model if for all positive f, g, ε there is a polynomial p such that for all $(T, \mathcal{M}) \in GM_{f,g}$ on set S of n sequences of length at least $p(n)$ generated on T, we have $\Pr[\Phi(S) = T] > 1 - \varepsilon$

Theoretical Comparison of Methods

- **Theorem 1** *[Warnow et al. 2001]*
  DCM_{NJ}+SQS is absolute fast converging for the GM model.
- **Theorem 2** *[Atteson 1999]*
  NJ is exponentially converging for the GM model.
- **Theorem 3** *[Szekely and Steel]* ML is exponentially converging for the GM model.

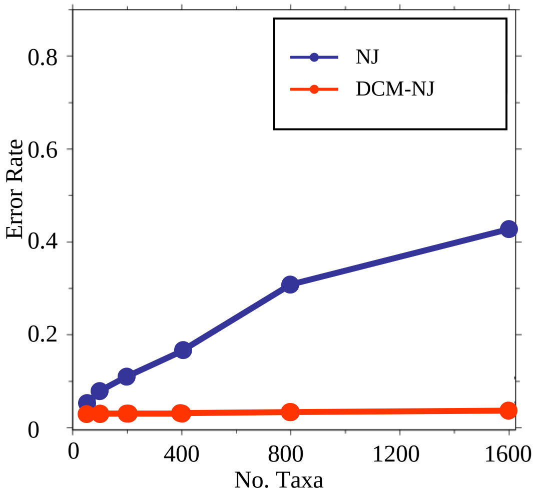
DCM-Boosting *[Warnow et al. 2001]*

- DCM+SQS is a two-phase procedure which reduces the sequence length requirement of methods.


Exponentially converging method
DCM
SQS
Absolute fast converging method

- DCM_{NJ}+SQS is the result of DCM-boosting NJ.

Main Result: DCM-boosting phylogenetic reconstruction methods *[Nakhleh et al. ISMB 2001]*

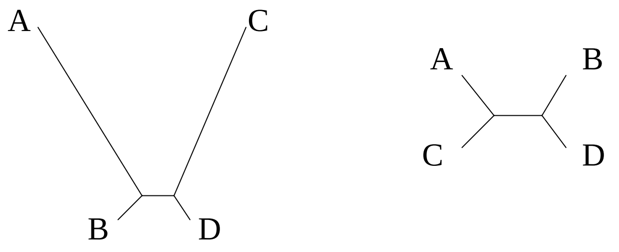


- DCM-boosting makes fast methods more accurate
- DCM-boosting speeds-up heuristics for hard optimization problems

Part III: Maximum Parsimony (30 minutes)

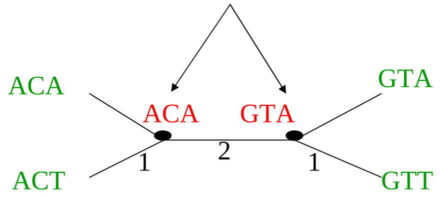
MP is not statistically consistent

- Jukes-Cantor evolution
- The Felsenstein zone



Maximum Parsimony: computational complexity

Optimal labeling can be
computed in linear time O(nk)



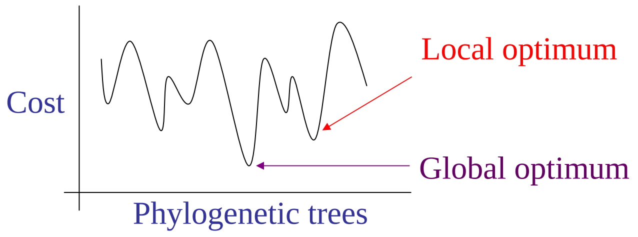
MP score = 4

Finding the optimal MP tree is **NP-hard**

Approximation algorithms

- 2-approximation algorithm: Compute MST on the graph where the vertex set is the set of sequences
- More generally, approximation algorithms for the Steiner Tree problem can be applied to the MP problem

Local search strategies



Heuristics for MP

- Hill-climbing based upon TBR, SPR, or NNI moves
- The Parsimony Ratchet
- Sectorial Search
- Disk-Covering

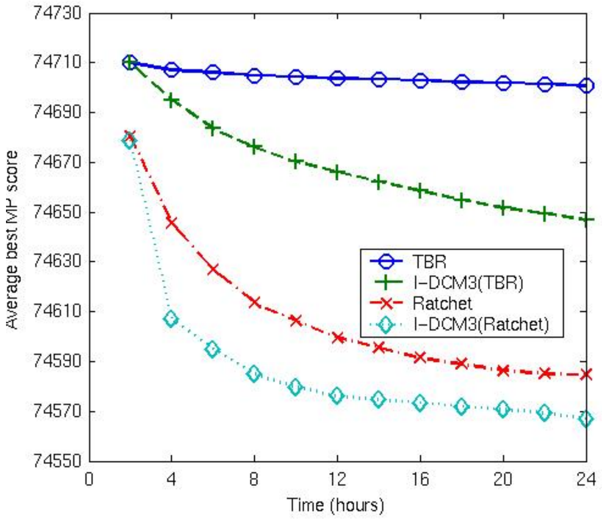
How good an MP analysis do we need?

- Our research (Moret, Roshan, Warnow, and Williams) shows that we need to get within **0.01%** of optimal MP scores (or better even, on large datasets) to return reasonable estimates of the true tree’s “topology”

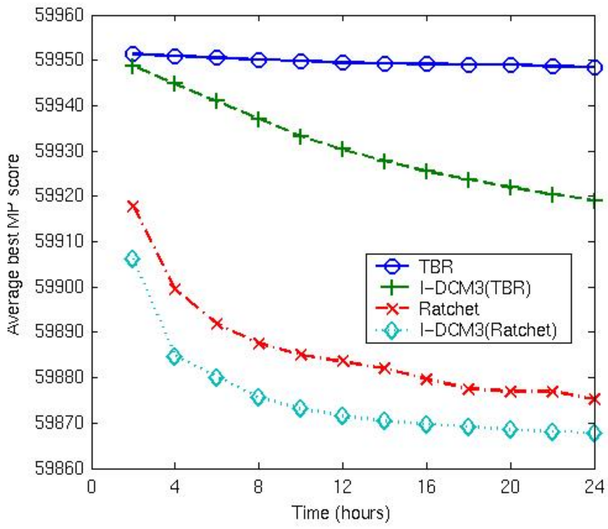
Comparison of MP heuristics

- **Methods:** TBR search, Ratchet, I-DCM3(TBR), I-DCM3(Ratchet)
- **Datasets:** Biological data
- **Experimental Methodology:**
  - On each dataset we ran 10 trials of each method (each trial for 24 hours).
  - We then plotted avg. best MP scores after fixed time intervals.
- **Implementation:** Ratchet was implemented using PAUP*4.0 and I-DCM3 was implemented by us using C++. We used Linux Pentium machines for our experiments.

2000 Eukaryotes sRNA (Gutell et. al.)



2594 rbcL DNA (Kallersjo et. al.)



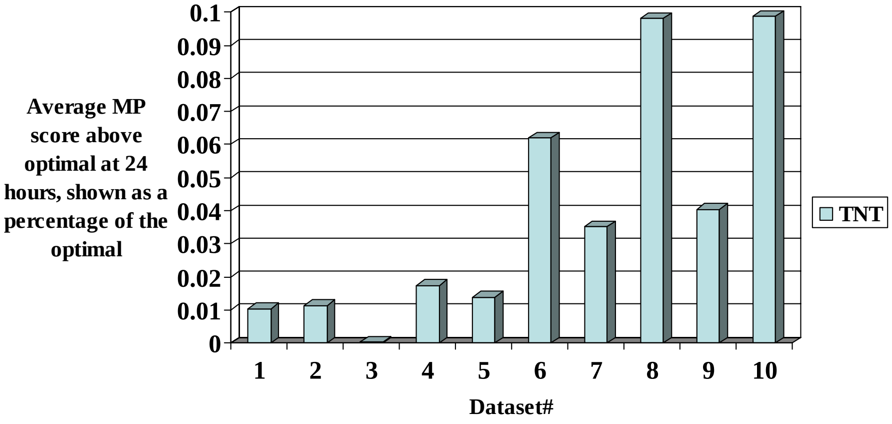
Datasets

Obtained from various researchers and online databases

- 1322 lsu rRNA of all organisms
- 2000 Eukaryotic rRNA
- 2594 rbcL DNA
- 4583 Actinobacteria 16s rRNA
- 6590 ssu rRNA of all Eukaryotes
- 7180 three-domain rRNA
- 7322 Firmicutes bacteria 16s rRNA
- 8506 three-domain+2org rRNA
- 11361 ssu rRNA of all Bacteria
- 13921 Proteobacteria 16s rRNA

Problems with current techniques for MP

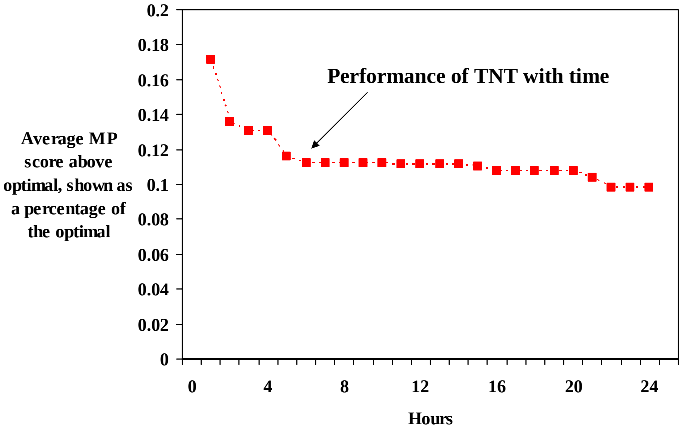
Average MP scores above optimal of best methods at 24 hours across 10 datasets



Best current techniques fail to reach 0.01% of optimal at the end of 24 hours, on large datasets

Problems with current techniques for MP

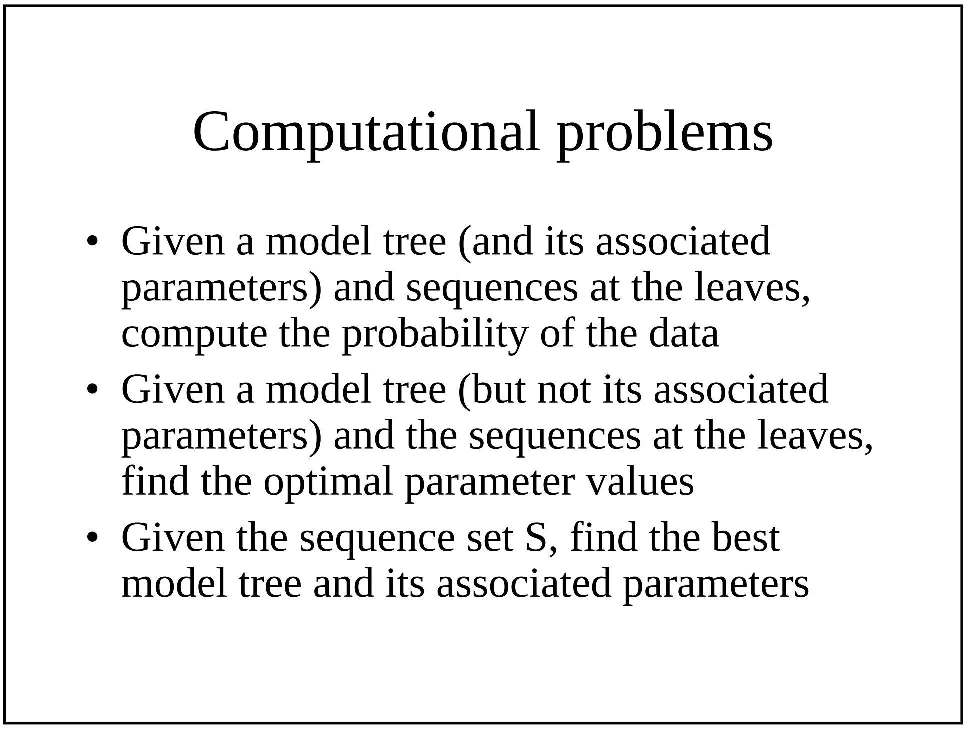
Best methods are a combination of simulated annealing, divide-and-conquer and genetic algorithms, as implemented in the software package TNT. However, they ***do not reach 0.01% of optimal on large datasets in 24 hours.***



Challenges

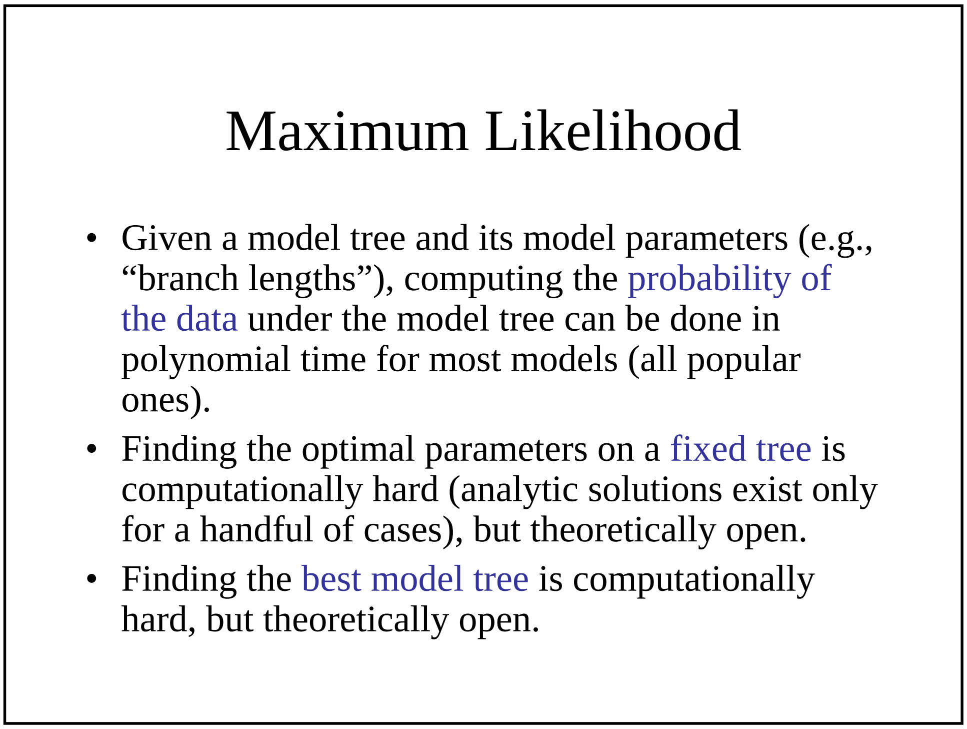
- Good lower bounds
- More effective heuristics
- Branch-and-bound
- Statistical performance issues

Part V: Maximum Likelihood (15 minutes)



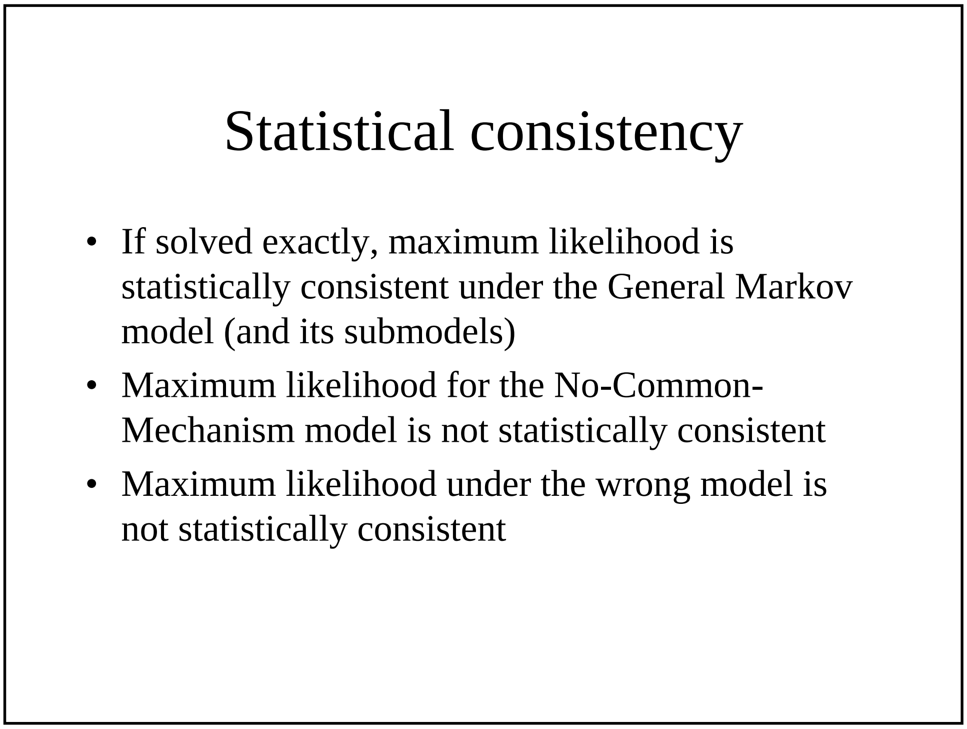
Computational problems

- Given a model tree (and its associated parameters) and sequences at the leaves, compute the probability of the data
- Given a model tree (but not its associated parameters) and the sequences at the leaves, find the optimal parameter values
- Given the sequence set S, find the best model tree and its associated parameters



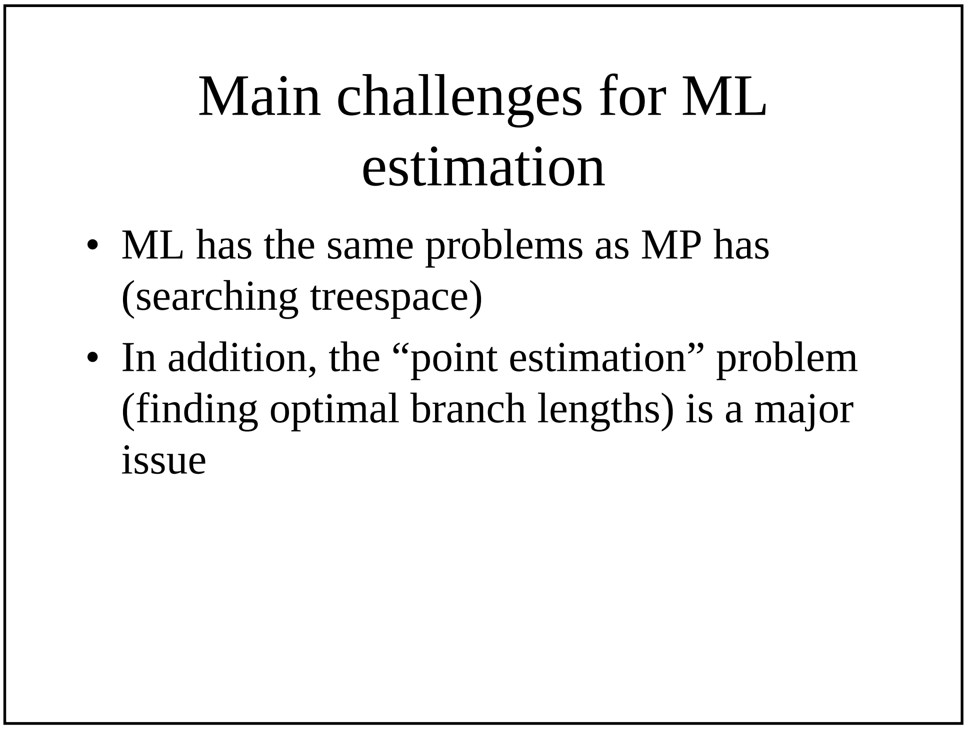
Maximum Likelihood

- Given a model tree and its model parameters (e.g., "branch lengths"), computing the probability of the data under the model tree can be done in polynomial time for most models (all popular ones).
- Finding the optimal parameters on a fixed tree is computationally hard (analytic solutions exist only for a handful of cases), but theoretically open.
- Finding the best model tree is computationally hard, but theoretically open.



Statistical consistency

- If solved exactly, maximum likelihood is statistically consistent under the General Markov model (and its submodels)
- Maximum likelihood for the No-Common-Mechanism model is not statistically consistent
- Maximum likelihood under the wrong model is not statistically consistent



Main challenges for ML estimation

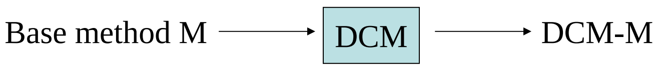
- ML has the same problems as MP has (searching treespace)
- In addition, the “point estimation” problem (finding optimal branch lengths) is a major issue

Part VI: Open problems/research directions (1 hour)

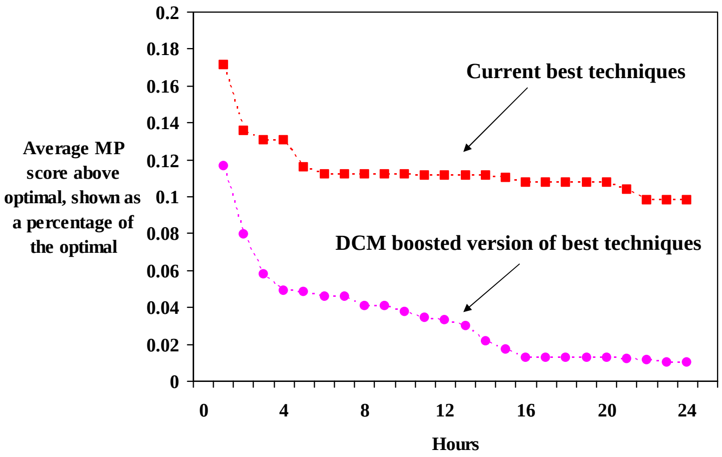
- Speeding up searches through tree-space
- Speeding up the ML evaluation of a fixed model tree topology (assigning branch lengths)
- Non-tree models
- New data (e.g., gene order and content)
- Supertree methods

"Boosting" MP heuristics

- We use "Disk-covering methods" (DCMs) to improve heuristic searches for MP and ML



Rec-I-DCM3 significantly improves performance

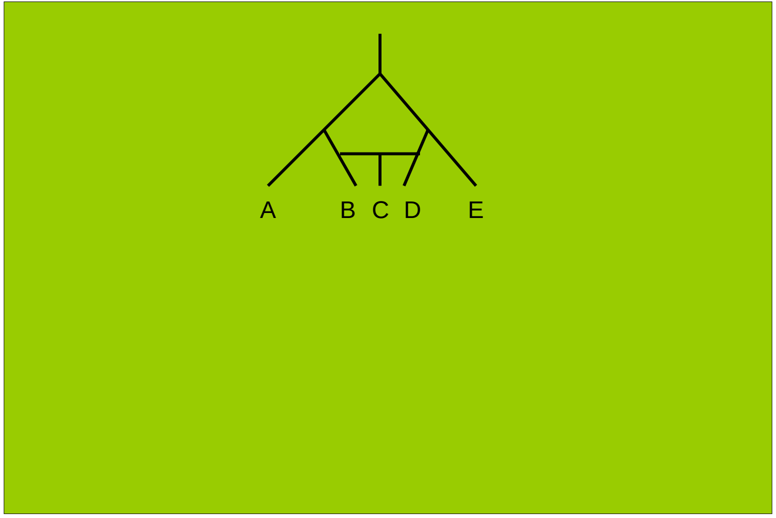


Comparison of TNT to Rec-I-DCM3(TNT) on one large dataset

Why Networks?

- Lateral gene transfer (LGT)
  - Ochman estimated that 755 of 4,288 ORF’s in E.coli were from at least 234 LGT events
- Hybridization
  - Estimates that as many as 30% of all plant lineages are the products of hybridization
  - Fish
  - Some frogs

Species Networks



Reconstructing Phylogenetic Networks

Main question: to combine, or not to combine?

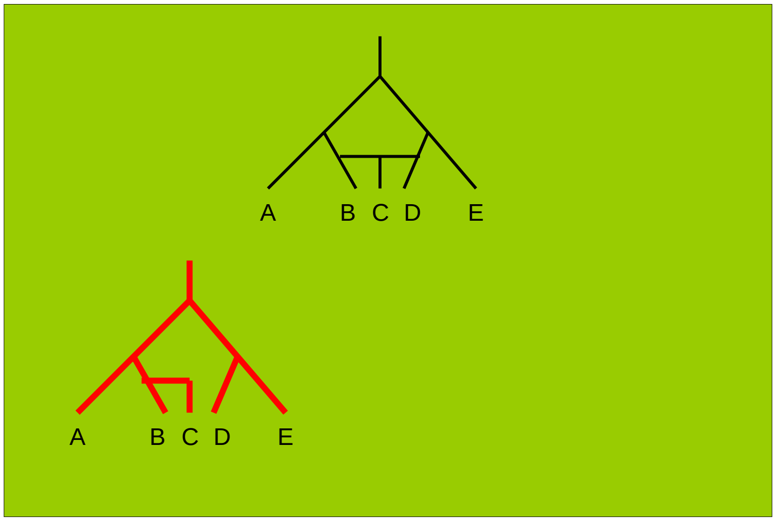
Separate analysis:

- Analyze individual genes separately
- Reconcile the resulting phylogenies

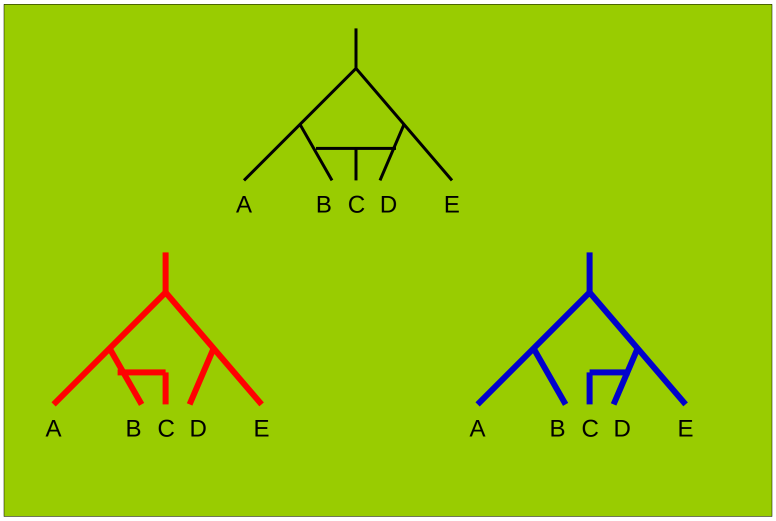
Combined analysis:

- Combine (via concatenation) the datasets, and attempt to infer the evolutionary history

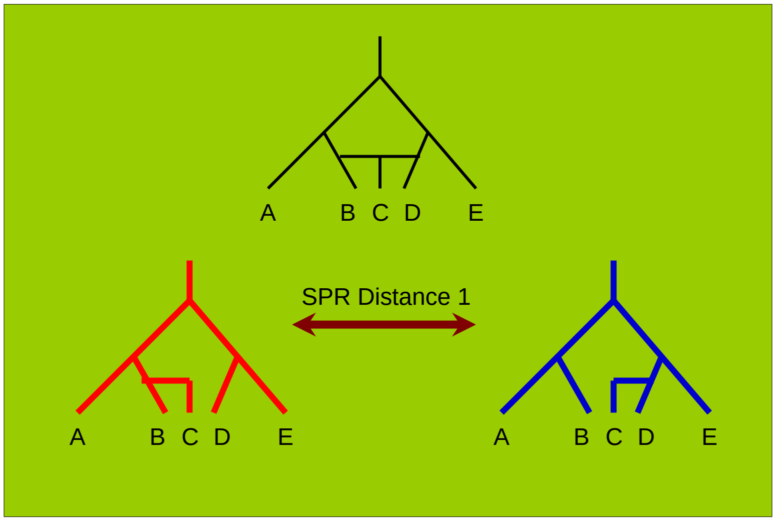
Gene Tree I in Species Networks



Gene Tree II in Species Networks



SPR Distances Among Gene Trees



Maddison’s Method

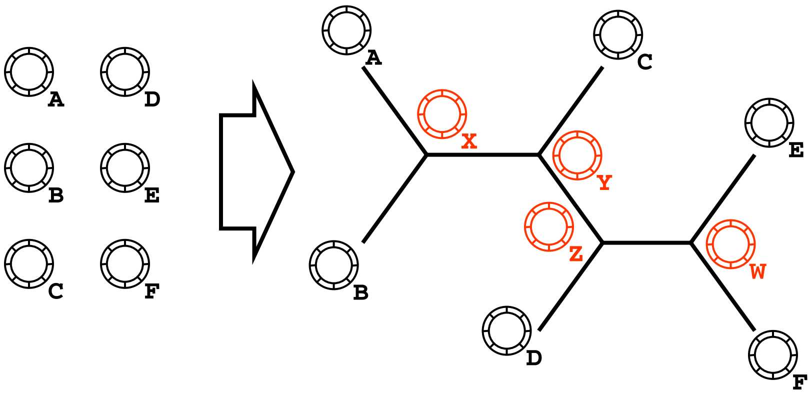
Given two gene datasets

- Construct two gene trees T1 and T2
- If SPR(T1,T2)=0
  - Return a tree
- If SPR(T1,T2)=1
  - Return a network with one reticulation event
- If SPR(T1,T2)>1, return FAIL

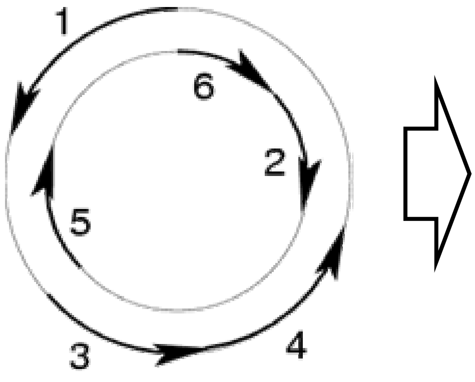
Open problems for reticulation

- Detecting reticulation
- Representing reticulate evolutionary scenarios
- Inferring reticulate evolution
- Visualization

Whole-Genome Phylogenetics

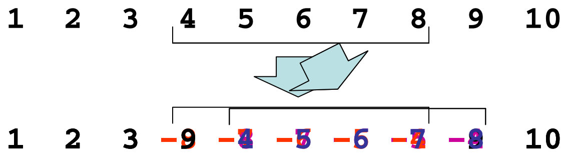


Genomes As Signed Permutations



1 –5 3 4 -2 -6
or
6 2 -4 –3 5 –1
etc.

Genomes Evolve by Rearrangements



- Inversion (Reversal)
- Transposition
- Inverted Transposition

Other types of events

- Duplications, Insertions, and Deletions (changes gene content)
- Fissions and Fusions (for genomes with more than one chromosome)

These events change the number of copies of each gene in each genome *("unequal gene content")*

Genome Rearrangement Has A Huge State Space

- DNA sequences : 4 states per site
- Signed circular genomes with n genes:

$2^{n-1}(n-1)!$ states, 1 site

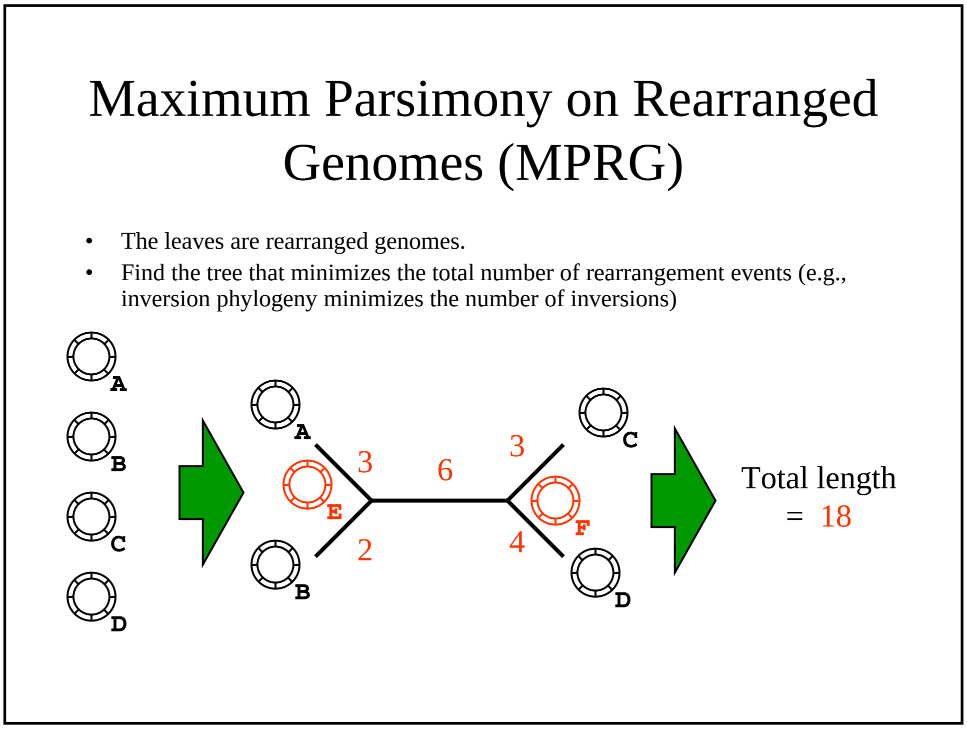
- Circular genomes (1 site)
  - with 37 genes (mitochondria): 2.56×10^{52} states
  - with 120 genes (chloroplasts): 3.70×10^{232} states

Why use gene orders?

- “Rare genomic changes”: huge state space and relative infrequency of events (compared to site substitutions) could make the inference of deep evolution easier, or more accurate.
- Our research shows this is true, but accurate analysis of gene order data is computationally very intensive!

Phylogeny reconstruction from gene orders

- Distance-based reconstruction: estimate pairwise distances, and apply methods like Neighbor-Joining or Weighbor
- **“Maximum Parsimony”:** find tree with the minimum length (inversions, transpositions, or other edit distances)
- Maximum Likelihood: find tree and parameters of evolution most likely to generate the observed data

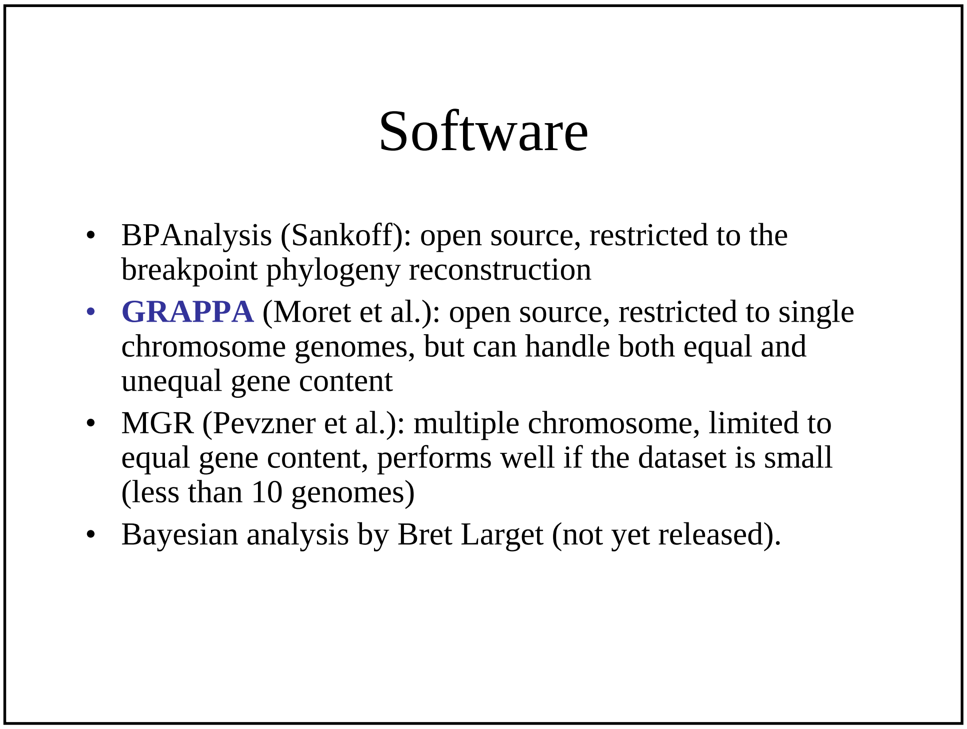


Maximum Parsimony on Rearranged Genomes (MPRG)

- The leaves are rearranged genomes.
- Find the tree that minimizes the total number of rearrangement events (e.g., inversion phylogeny minimizes the number of inversions)

A
B
C
D
A
3
6
3
C
E
F
2
4
B
D

Total length = 18



Software

- BPAnalysis (Sankoff): open source, restricted to the breakpoint phylogeny reconstruction
- **GRAPPA** (Moret et al.): open source, restricted to single chromosome genomes, but can handle both equal and unequal gene content
- MGR (Pevzner et al.): multiple chromosome, limited to equal gene content, performs well if the dataset is small (less than 10 genomes)
- Bayesian analysis by Bret Larget (not yet released).